LOD 技术在德国图书档案馆的应用

董杰

哈尔滨商业大学图书馆 哈尔滨 150028

摘要：[目的/意义]关联开放数据(LOD)已广泛应用于很多产业、非营利性组织和政府。图书档案馆是 LOD 技术的早期使用者之一，这也促进了 LOD 技术的发展，德国是图书档案馆业非常发达的国家，有很多 LOD 应用于图书档案馆中的成功案例。[方法/过程]采用文献调研、网络调查、内容分析法，分析 LOD 技术在德国图书档案馆中成功应用的案例。[结果/结论]案例揭示了在计算机科学领域，如人工智能、数据库和图书档案馆研究课题之间的关系。总结了德国的实践经验，为我国发展相关的实践提供更多的参考。

关键词：关联开放数据；LOD；德国；图书馆；档案馆；应用

分类号：G250

基金项目：本文系国家自然科学基金“基于吸附机理及膜污染控制的吸附/膜法藻源型有机质去除的调控策略研究”(项目编号：51408169)、哈尔滨商业大学博士科研启动项目(项目编号：14LG15)和黑龙江省自然科学基金项目“基于多平面管理的黑龙江省高校人力资源管理模型及示范系统研究”(项目编号：F201217)研究成果之一。

作者简介：董杰（ORCID：0000-0001-5758-0139）馆员，博士，E-mail：124348423@qq.com。

1 引言

德国有 8 000 多家公立图书档案馆，其中约一半为州立、市立图书档案馆，一半为教会图书档案馆，还有私立图书档案馆 10 000 多家，平均约 4 000 多人就有一家图书档案馆。可见，德国是图书档案馆业发达国家之一[1]。

越来越多的国家和国际组织更加重视数字图书档案馆之间的合作。越来越多的用户将数据发布到网络上，形成了全球性的数据网络(Web of Data)。与文档网络相比，结构化的数据网络形成了更加复杂的关系网，更容易检索 Web 数据，人和机器也更容易理解这些数据。2017 年 2 月[2]，W3C 项目发布了新的关联开放数据云图(Linked Open Data Cloud, LOD Cloud)，见图 1，建立了新的视觉模型，开放关联数据集的数量增长了数十倍达到了几百个，内容包含了出版物、跨领域、媒体、语言学、地理、用户生成内容、政府、环境、生命科学和社交网络等多个领域。LOD 将多个领域关联开放数据资源集成为一个可视化的互联网络。从情报学的角度分析，这是在引证、合著等知识网络后的新的网络型态[3]。

近年来，数字图书档案馆进一步促进信息资源共享，而数字图书档案馆面临的问题是如何提供对大量数据访问的服务，这些数据是隐藏的、不可访问的，并且存储在数据竖井中。随着 Web 对异构数据访问技术的发展，LOD 可以实现对元数据的发布，这将使图书档案馆的馆藏资源能够以可持续的方式被搜索、链接

和访问[4]。另一方面，LOD 是运用语义技术发布和共享信息的最佳方法，并且可以访问大量的异构数据，这可以激发更多应用程序的开发。LOD 可以帮助数字图书档案馆摆脱数据竖井，将其数据发布成为结构化数据；并为图书档案馆带来很多应用价值[2]。

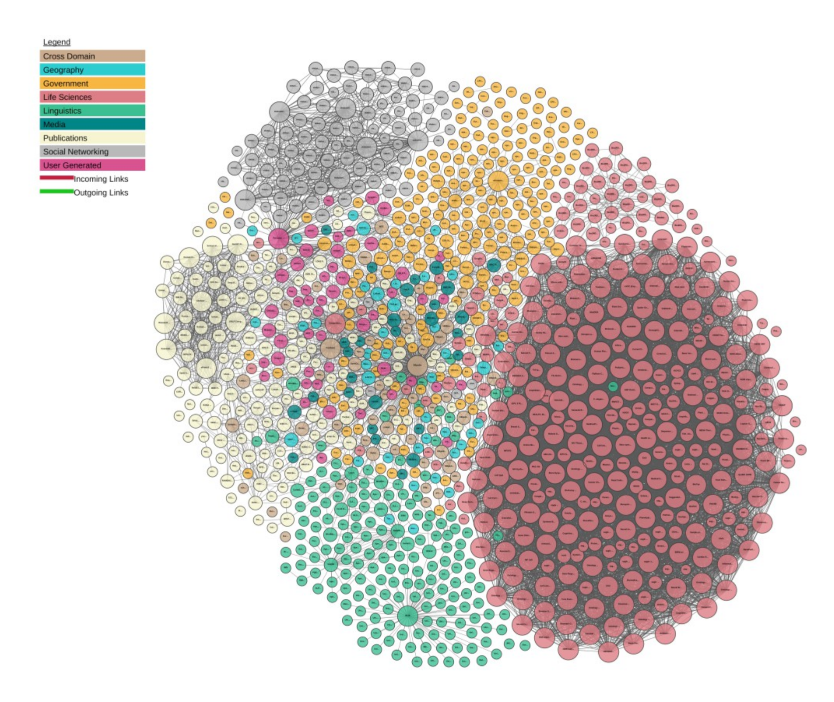


图 1 关联开放数据云图

2 德国数字图书档案馆的成功案例

德国数字图书档案馆的成功案例描述了数字图书档案馆在信息供应方面的不同需求，并总结了相关数据技术是如何满足这些需求的。此外，明确了 LOD 技术在数字图书档案馆应用中的主要优势。

2.1 关联数据价值链的成功应用

德国数字图书档案馆的研究项目将公开可用数据转换为关联数据。绝大多数的数据都是由研究机构产生的。将关联数据价值链（见图 2）引入到商业工程师的模型中，可以使成功商业案例概念化，确定角色的分配、组合和参与，但所选择的数据及其转换过程可能存在固有的风险，例如：使用权限、隐私策略、数据可用性和角色激励、数据质量和可信度、数据来源、透明数据转换和互连等。

德国莱布尼茨经济信息中心（Leibniz Information Centre for Economics，ZBW）将关联数据价值链应用到 BBC[3] 的现有业务案例中，并在此过程中对潜在的风险进行了测试。总的来说，关联数据价值链有助于识别和分类潜在的风险，这些风险可由相应的工程师来处理，而且还建立了能清晰了解完整关联数据生成周期的方法。这个模型易于在其他学科中应用，如数字图书档案馆、生命科学和媒体等，有助于关联数据的发

布，并可指出可能出现的潜在问题，这些问题可能出现在数据转换和数据间链接过程中[5]。

图 2 关联数据价值链模型[5]

2.2 LOD 技术在数字期刊中检索作者信息的应用

LOD 技术在数字期刊中的应用价值之一是可以通过关联数据来实现现实世界作者与数字期刊中作者的联系。在 ZBW 数字环境分析系统中，在处理与人有关的信息时面临的问题是作者姓名识别和消除歧义。分析系统在个人资料中找到相关的个人信息，如专业知识，对社交媒体的影响以及出版物的数量等。基于 LOD 的分析系统可以在组织和机构的人员分配等方面发挥至关重要的作用。因此，找到有关作者的正确信息对于提高数字期刊的整体可见性和效率至关重要[6]。

在 LOD 的基础上，德国科学家开发了 CAF-SIAL 平台，见图 3，可以搜索并提供来自关联数据人员信息（http://cafsial.lod-mania.com）。CAF-SIAL 平台运用一组启发式技术，将一个人的相关信息从 DBpedia 中识别出来，通过对“URI”技术应用一个“关键字”来提取。这个提取的信息被进一步过滤，并集成到一个概念聚合框架下，这个框架随后被呈现为一个概要文件[7]。

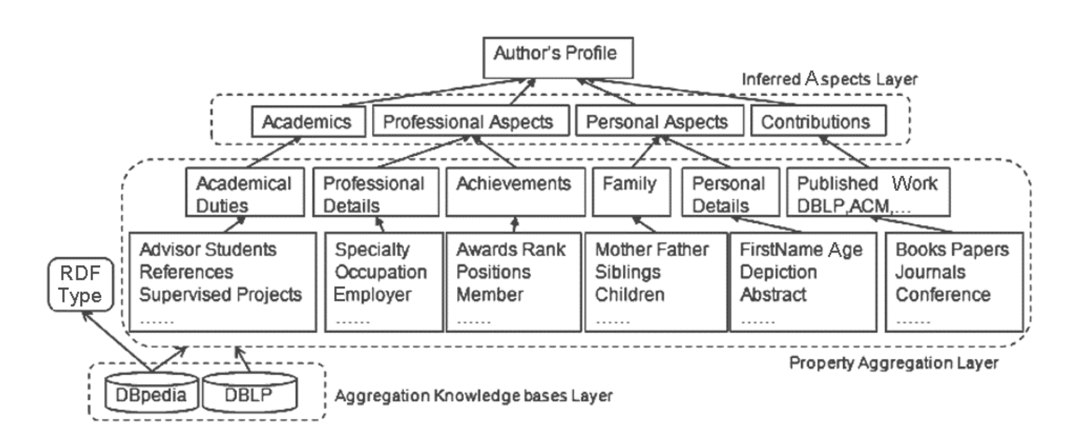


图 3 作者信息集成模型[7]

在图书档案馆环境中，DBpedia 和 DBLP 体现了应用程序的实用性，进一步扩展了数字期刊作者与 LOD 的相关语义资源之间的联系。通过该应用程序能够识别、消除歧义，检索和构造有关来自这些数据集的作者的相关信息。该系统构建了一个全面的作者资料库，可以提供作者信息(个人和专业信息)，并列出他的学术成果（http://dblp.l3s.de/d2r/）。

这类系统可以应用在更广泛的学术交流领域中。搜索的主体可以扩展到集成的权限文件，如德国国家图书档案馆的综合授权文件(GND)（http://www.dnb.de/EN/gnd）和虚拟国际权威文件(VIAF)（https://viaf.org/），以获得更多完整的结果。权限文件所包含的关键词和描述符在编目过程中被分配给一个出版物，这样可以进一步简化搜索和检索过程。

2.3 LOD 技术在关联数据发布的应用

在过去的几年里，LOD 对的数据的开放起到了重大作用，并已成为最重要的类库应用程序之一。这些存储库是用于收集、发布、传播和存档数字科学内容的系统。在数字图书档案馆的应用方面，EconStor 可以使存储库中的科学论文的元数据以机器可读的方式提供给读者（http://econstor.eu）。EconStor 是德国国家经济图书档案馆的开放访问服务器，为出版经济学研究论文提供了平台。EconStor 目前提供近 100 个机构的科学论文以及超过 8 万份完整的文本文件的全文访问[8]。

D_2RQ 框架可以将关系数据集转换为可理解的语句，并将 EconStor 存储库数据发布为关联数据（http://d2rq.org/）（见图 4），步骤如下：第一步，将开放存储库作为关系数据库；第二步，通过使用词汇表，将出版物和作者映射到 D_2R 服务器转换为映射文件；最后，存储库数据通过使用 D_2R 服务器进行转换，并将其作为关联数据和 SPARQL 端点进行查询（http://linkeddata.econstor.eu/beta/snorql/.）。存储库的内容可以直接作为关联开放数据发布，并且能够关联到有价值的外部数据集，从而使存储库中的数据能够上下文关联并有意义。通过将 EconStor 作为关联数据库发布实现了以下预期目标：通过将科学论文发表在语义网上，从而使当前研究成果能够出版和传播；成功地使典型的存储库系统(如 DSpace)转变成语义 Web 开

放内容，并将其集成到关联数据流中；通过 SPARQL 查询模式，使查询分布式的研究信息成为可能，如可以查询 2012 年之后由欧洲研究机构出版的所有关于金融危机的文章。

将 EconStor 作为关联数据发布，对 mashup 应用程序（这些应用程序可以从不同的相关关联数据存储中对数据进行管理）的开发带来了潜在的影响。从软件工程的角度来看，该研究提供了将存储库的内容发布为关联开放数据的方法。因此，图书档案馆员、仓库管理员和软件开发人员对此都产生了极大的兴趣。

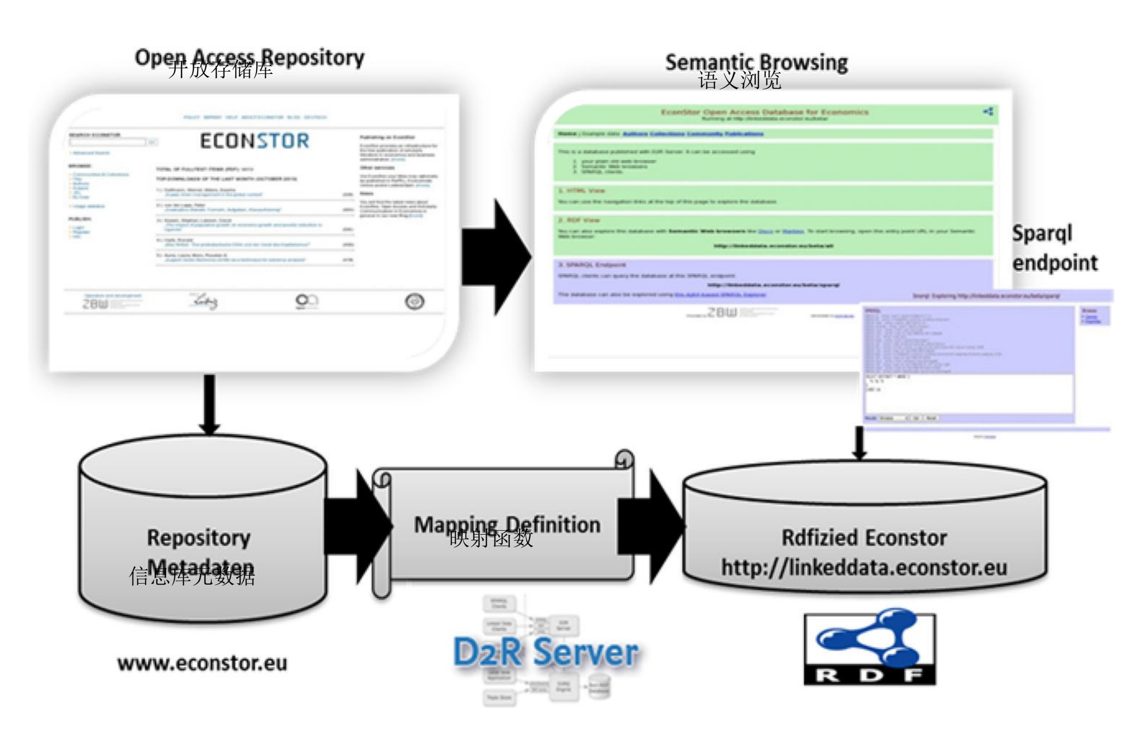


图 4 EconStor 关联数据发布模型[8]

3　图书档案馆科学中 LOD 的研究

3.1 实体解析

“实体解析”指的是识别两种关联开放数据中的资源是否指向同一个真实世界中的实体。这是一项具有挑战性的任务，因为资源没有自己的身份，其意义仅通过语义描述和连接资源的属性来定义，解决这个问题的一种方法是通过手动调整。德国国家图书档案馆的综合管理局文件包含与 DBpedia 等相关的作者信息[9]。然而，手动调整非常耗费人力，并且不可能实现大型数据集的合并。如 DBpedia 数据库中包含 364 000 个数据，德国国家图书管理局数据库中包含 1 797 911 个数据，国会图书馆数据库中包含 3 800 000 个数据，虚拟国际权威档案（VIAF）大约有 1 000 万个数据（VIAF 组合了不同国家图书档案馆的多个名称权限文件），这些数据库都非常庞大， 因此，仅通过姓名、合作者、职称和地点对其进行实体解析通常是不够的

[10]。

3.2 模式匹配

模式匹配与实体解析所面临的问题相似。链接开放数据的目标是通过参考其他现有词汇的概念来定义和发布自有词汇。然而，不同词汇的整合以及他们所描述的数据都是很重要的，即使是具有类似模式的数据库也是如此。在运用模式集成来改进图书档案馆服务的过程中对模式匹配质量要求是非常高的[11]。因此，通过人工调整叙词表的方法来对不同作品进行模式匹配。如 ZBW 对经济学词典 STW（http://zbw.eu/stw/versions/latest/about）与其他词典（如社会科学中的 TheSoz，http://lod.gesis.org/pubby/page/thesoz/）在 2004-2005 年期间手动创建了数千个映射。为了描述映射，关键字之间的关系通常用简单知识组织系统（SKOS）词汇来描述（http://www.w3.org/2004/02/skos/）。由于叙词表通常有几千甚至一万个主题词和相应的同义词，需要用自动的方法进行模式匹配，因此，2012 年 ZBW 启动了比对评估计划（OAEI）。OAEI 旨在比较不同的模式匹配技术，并就本体匹配方法的评估达成共识（http://oaei.ontologymatching.org/.）。

3.3 分布式数据管理

LOD 数据是是分布式数据，其中 VIAF 是一个很好的例子，其中有十几个国际组织合作构建分布式图书档案馆资源网络，不仅有出版商，还包括个人和组织。为了访问分布的数据，需要应用联合查询技术，并且搜索出数据源信息及信息存储形式。

在语义 Web 中，研究人员已经开发了各种不同的技术，如用于关联打开分布式数据的查询技术、用于对关联开放数据进行流处理的技术以及用于搜索服务数据和数据源的技术。然而，到目前为止，还不清楚哪种方法最适合访问分布式数据[12]。

此外，在提供图书档案馆搜索服务时，还需要考虑搜索结果排名，以便满足用户的查找需求。像网络搜索一样，用户也认为搜索结果中第一个链接比其他链接更重要或更相关。为了应对这一问题，ZBW 的 DFG（German Research Foundation，德国研究基金会）项目开发的 LibRank 实现了这一目标（http://www.librank.info/）。

3.4 自动索引

与数据库社区的索引概念相反，在图书档案馆中，索引是指为科学出版物、档案等文件分类标出多个标签。索引的一种方法是手工标记，德国科学家使用 STW 标记了超过 160 万份经济学出版物。这些出版物每篇平均标注了 5 个 STW 主题词。另外，运用发布服务器 EconStor 实现了 STW 和其他叙词表的作者和关键词的自动发布。

此外，德国国家图书档案馆每年出版的电子出版物数量显著增加，需要采用自动化的索引文献方法。

为此开发了用于 PDF 分类的自动化方法。如德国国家图书档案馆的 PETRUS 项目使用支持向量机对 100 个类别（Sach-gruppen）进行分类。DFG 资助的项目 GERHARD 在 20 世纪 90 年代研究了自动索引科学 Web 内容的方法。

研究人员运用十进制分类法（UDC）将约 100 万个文档自动编入索引中。UDC 索引使用 3 种语言（德语、英语、法语）。使用 Oracle 关系数据库管理系统可以进行全文索引（ConText）。科学文献的自动化索引迄今为止仍然是非常活跃的研究领域[10]。

在最近的 ZBW 项目中正在进行应用关联开放数据自动索引科学文档的工作。运用 kNN 分类器、实体检测和 HITS 算法来评估 STW 对特定文档的匹配性。ZBW 开发应用自动分度实验的优点是不需要昂贵的培训[13]。

虽然多数人认为术语“自动索引”过程中是没有人的参与的，但上述技术需要人为干预才能准确运行 。事实上，在运行过程中需要图书档案专业人员运用专业知识不断监测自动索引主题词的质量，使其能正确反映主题。

3.5 索引非文本内容

除了 PDF 格式的科学出版物和图书档案馆索引的网站等文字内容外，还有大量的非文字内容，如社交媒体和视听材料。这些材料包括传统科学内容的映射、社会媒体资料，还有研究数据， ZBW 在欧盟项目 EEXCESS 中解决了这些非文本内容的索引问题（http://eexcess.eu/）。这个想法是将结构化科学内容（元数据、全文本、段落、引文和其他内容）与社交媒体渠道中的非正式和临时内容进行自动结合，以便关联主题、对象（文本和非文本资源）以及用户。在实体解析、多种模式索引以及跨媒体检索内容方面也存在了一些问题。

为了解决多模式检索的问题，ZBW 开发了一种新渠道，以便更好地理解包含在科学出版物中的图表。该渠道通过不同方法（如数据挖掘和计算机视觉等技术的组合）从图表中自动提取多项文本信息。这允许对信息图表进行文本搜索，并将其与科学出版物的文本内容相结合[14]。

3.6 数据出处

虚拟国际权威文件（Virtual International Authority File，VIAF）可以使书目记录在跨组织、跨境、跨语言中检索。通过匹配和链接开放权限的文件可以降低成本并增加授权文件的实用性。然而，在跨境、跨语言的情况下，出现了新的问题：如何跟踪数据/元数据（重新）使用？图书档案馆 A 使用图书档案馆 B 的（部分）记录时如何参考元数据？如何评估合并到系统中的数据/元数据的可信度？

为了解决跟踪数据来源的问题，图书档案科学界开发了用于描述图书档案馆资源的复杂模型。FRBR 模型可以描述同一图书档案馆资源的不同变体，如同一本书的不同印刷本，或不同的语言翻译版本

（http://www.ifla.org/publications/functional-requirements-for-bib liographic-records）。因此，它不仅适用于书籍，也适用于任何资源。另外，RDA 模型可以描述任何种类的内容，包括在线媒体。RDA 还允许将信息来源附加到不同的数据上（http://www.rda-jsc.org/rda.html）。Europeana 数据模型可以查询创建元数据记录的人员和资源本身的来源（http://www.europeana.eu/portal/）。

然而，仍然缺少一种能可靠验证元数据来源的方法。由 A. Kasten 等人开发的数字签名图形数据的框架可以用来跟踪元数据的来源。它用数字签名来标记图形并将数据与网络上的签名一起发布，例如关联打开数据。这可以跟踪元数据的来源，建立一个“信任网络”[15]。

此外，像语义搜索引擎 Sig.ma 这样的应用程序能够为 LOD 的实体搜索提供支持，并根据来源提供过滤结果。不幸地是，该项目已经终止[16]。

表 1 总结了 LOD 技术在德国数字图书档案馆的具体应用及其缺陷，揭示了在图书档案馆研究领域 LOD 技术进一步的研究方向。

表 1 图书档案馆科学中 LOD 的研究比较

序号	内容	具体应用	缺陷
1	实体解析	识别关联开放数据的两种资源是否指向同一个真实的世界实体的问题	需要人工调整，非常昂贵，不能合并大型数据集
2	模式匹配	定义和发布自我描述的词汇，通过模式集成来改进图书档案馆服务	需要人工调整叙词表模式匹配不同作品
3	分布式数据管理	用于关联打开或查询在网络上高度分布数据	不清楚访问分布式数据的最适方法，需要考虑结果排名
4	自动索引	自动索引科学文档	需要图书档案科学家不断监测自动建议的描述符的质量
5	索引非文本内容	解决了传统科学内容的映射、社会媒体资料、研究数据等非文本内容的索引	实体解析、跨媒体检索内容还需要解决
6	数据出处	FRBR 概念被并入到 RDA 中，以描述任何种类的内容。Europeana 数据模型可预测创建元数据记录的人员和资源本身的来源	缺少感知信息的来源应用程序

4 德国成功经验对我国的启示

数字化信息的收集、储存、应用及长久保存等诸多问题与数字技术与网络技术的发展密不可分。因此，德国图书档案馆从 1998 年起参加了欧盟创建的 “欧洲网络化缴存图书馆”等多个项目的工作，主要研究数字资源保存和应用等技术问题，构建基础的网络平台，开发多媒体传输技术等系统，研究迁移和仿真信息再现技术等。至今，德国图书档案馆基于 LOD 技术开发出的很多技术都具有普适性和应用性。其中一些技术甚至为世界数字图书档案馆的发展做出了积极的贡献。推行科学技术精神的德国品质也在图书档案馆

的技术领域表现出来，其 LOD 技术在图书档案馆的应用在国际上也具有极其重要的地位。

随着 LOD 中数据集的快速增长， LOD 技术在图书档案馆信息服务中的应用也越来越广泛。LOD 在我国的图书档案馆应用中还存在着一些不足，一些研究还局限于理论层面，没有真正地成为我国图书档案馆中可操作的应用技术，而这些技术可为将来的数字图书档案馆应用提供基本技术支持，且应用广泛。通过对基于 LOD 技术在德国图书档案馆的应用的比较（见表 1），可为图书档案馆中的很多实践工作指明进一步的研究方向。在我国，将 LOD 技术引入图书档案馆已经迫在眉睫，通过学习德国的经验，基于已有的条件搭建基于 LOD 的关联应用平台，在实践中应用已有的方法和工具解决相关问题。图书档案馆利用这些新技术将会产生新的服务。

参考文献

[1] 王永丹. 德国公共图书馆服务初探[J]. 图书馆理论与实践, 2016(2): 8-11.

[2] BERNERS-LEE T. Linked-data design issues. W3C design issue document[EB/OL]. [2017-01-20]. http://www.w3.org/DesignIssue/LinkedData.html.

[3] 夏立新, 谭荧. LOD 的网络结构分析与可视化[J]. 现代图书情报技术, 2016(1): 65-72.

[4] HEATH T, BIZER C. Linked data: evolving the web into a global data space[M]//Synthesis Lectures on the Semantic Web: theory and technology. San Rafael: Morgan and Claypool, 2011: 1-136.

[5] LATIF A, SAEED A U, HO¨FLER P, et al. The linked data value chain: a lightweight model for business engineers[C]// 5th international conference on semantic systems. Graz: Graz Technical University Press, 2009: 568-575.

[6] LATIF A, AFZAL M T, HELIC D, et al. Discovery and construction of authors' profile from linked data (a case study for open digital journal) [C]//CEUR workshop proceedings. Raleigh: LDOW, 2010: 628.

[7] LATIF A, AFZAL M T, HOFLER P, et al. Turning keywords into URIs: simplified user interfaces for exploring linked data[C]// Proceedings of the 2nd international conference on interaction sciences: information technology, culture and human. Seoul: Int. Conf. Interaction Sciences, 2009: 76–81.

[8] LATIF A, BORST T, TOCHTERMANN K. Exposing data from an open access repository for economics as linked data[J]. D-Lib magazine, 2014, 20(9): 9-10.

[9] HALPIN H, PRESUTTI V. An ontology of resources: solving the identity crisis[C]//European semantic Web conference. Heraklion: Lecture notes in computer science, 2009: 521–534.

[10] NEUBERT J, TOCHTERMANN K. Linked library data: offering a backbone for the semantic web[C]// Third knowledge technology week. Kajang: CCIS, 2011: 37–45.

[11] WICK M L, ROHANIMANESH K, SCHULTZ K, et al. A unified approach for schema matching, coreference and canonicalization[C]//Proceeding of the 14th ACM SIGKDD, international conference on knowledge discovery and data mining. New York: ACM, 2008: 722–730.

[12] KONRATH M, GOTTRON T, STAAB S, et al. Schemex—efficient construction of a data catalogue by stream-based indexing of linked data[J]. Journal of Web semantics: preprint server, 2012(16): 52-58.

[13] PETERS I, SCHERP A, TOCHTERMANN K. Science 2.0 and libraries: convergence of two sides of the same coin at ZBW Leibniz Information Centre for Economics[J]. IEEE STC social networking, 2015, 3(1): 149-157.

[14] BOSCHEN F, SCHERP A. Multi-oriented text extraction from information graphics[C]// Symposium on document engineering (DocEng). Lausanne: ACM, 2015.

[15] KASTEN A, SCHERP A, SCHAUB P. A framework for iterative signing of graph data on the web[C]// The semantic Web: trends

and challenges proceedings. ESWC 2014. Lecture Notes in Computer Science. Anissaras: Springer, 2014: 146–160.
[16] TUMMARELLO G, CYGANIAK R, CATASTA M, et al. Sig.ma: live views on the Web of data[J]. Web Semantics, 2010, 8(4): 355–364.

Application of LOD Technology in German Libraries and Archives

Dong Jie
Library of Harbin University of Commerce, Harbin 150028

Abstract: [Purpose/significance] Linked Open Data (LOD) has been widely used in large industries, as well as non-profit organizations and government organizations. Libraries and archives are ones of the early adopters of LOD technology. Libraries and archives promote the development of LOD. Germany is one of the developed countries in the libraries and archives industry, and there are many successful cases about the application of LOD in the libraries and archives. **[Method/process]** This paper analyzed the successful application of LOD technology in German libraries and archives by using the methods of document investigation, network survey and content analysis. **[Result/conclusion]** These cases reveal in the traditional field of computer science the relationship among research topics related to libraries and archives such as artificial intelligence, database and knowledge discovery. Summing up the characteristics and experience of German practice can provide more reference value for the development of relevant practice in China.

Keywords: Linked Open Data; LOD; Germany; library; archives; application